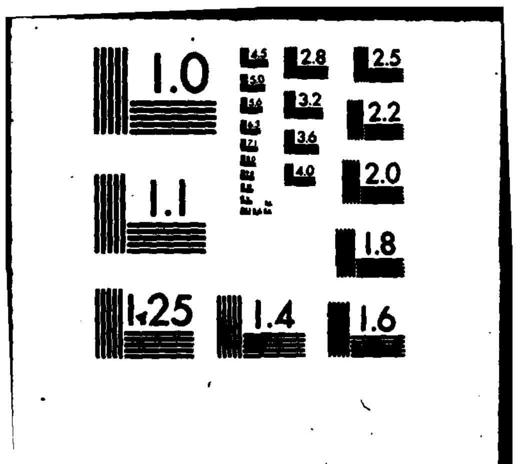
AD-A081 490 PURDUE UNIV LAFAYETTE IND DEPT OF STATISTICS F/G 12/1
SOME STATISTICAL TECHNIQUES FOR CLIMATOLOGICAL DATA.(U)
JAN 80 S S GUPTA, S PANCHAPAKESAN N00014-75-C-0455
UNCLASSIFIED MMS-80-1 NL

END
DATE
FILMED
4-80



MICROCOPY RESOLUTION TEST CHART

PURDUE UNIVERSITY

DEPARTMENT OF STATISTICS

DIVISION OF MATHEMATICAL SCIENCES

2

DTIC
ELECTE
MAR 3 1980
S C D

6

SOME STATISTICAL TECHNIQUES FOR CLIMATOLOGICAL DATA*

10 Shanti S. Gupta a[redacted]d S. Panchapakesan
Purdue University and Southern Illiniois University

14 NMS-80-1

Department of Statistics
Division of Mathematical Sciences

9 Mimeograph Series #80-1 ✓

11 Janu[redacted]80

12 13

* This research was supported by the Office of Naval Research Contract
15 N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

391 730

1B

SOME STATISTICAL TECHNIQUES

FOR CLIMATOLOGICAL DATA*

Shanti S. Gupta and S. Panchapakesan

Purdue University and Southern Illiniois University

ABSTRACT

Statistical methods are increasingly being applied in the analysis of climatological data. A brief introduction to subset selection approach in multiple decision theory is given to illustrate the potential applications in climatology.

1. Introduction.

The need for statistical methodology in analyzing data that arise in meterology and climatology has long been recognized. Satisfactory statistical models have been found to describe data relating to precipitation; see for example, Crutcher (1968) and Mielke (1973). Time series data occur commonly in climatological studies. Some of the important and interesting problems arise in connection with weather modification experiments, objective weather forecasting and classification of meterological patterns. Some relevant references are Braham (1979), Bradley, Srivastava and Lanzdorf (1979), Lund (1971), McCutchan and Schroeder (1973), Mielke (1979), Neyman (1977, 1979), and Neyman, Scott and Wells (1969) (see also the bibliography by Hanson et al (1979)). In the present paper, our main interest is in two types of problems. The first deals with comparisons of sites (weather stations) based on appropriate characteristics of weather data. For example, we may compare these locations on the basis of mean annual temperature or the variability of temperature during the year. The second problem relates to selection of the best

* This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

predictor variables in the regression model for prediction. We discuss ranking and selection formulation of these multiple decision problems.

Section 2 deals with the basic formulations of the ranking and selection problems. Some specific subset selection procedures are briefly described in Section 3. These deal with selection from normal populations in terms of the means, from gamma populations in terms of the scale parameters, and from multivariate normal populations in terms of multiple correlation coefficients. The next section is concerned with selection of the best set of predictor variables in a regression model.

2. Ranking and Selection Theory - Basic Formulations.

To describe the formulation of ranking and selection problems, let us consider k independent populations $\pi_1, \pi_2, \ldots, \pi_k$ where π_i is characterized by the distribution function $F(x, \theta_i)$ where θ_i is a parameter which represents the 'worth' of the population. For example, θ_i may be the weather characteristic of the ith location. Let $\theta_{[1]} \leq \cdots \leq \theta_{[k]}$ denote the ordered θ_i. To be specific, let us say π_i is preferable to π_j if $\theta_i > \theta_j$ so that the best population is the one associated with the largest θ_i. Ranking and selection problems have been generally formulated using either <u>indifference zone approach</u> or the <u>subset selection approach</u>. Under the indifference zone formulation of Bechhofer (1954), we want a procedure R which will select the best population with a minimum guaranteed probability P^* $(1/k < P^* < 1)$ whenever $\theta_{[k]} - \theta_{[k-1]} \geq \theta^*$ where $\theta^* > 0$ and P^* are specified in advance. The problem is to determine the minimum sample size needed in order to meet this requirement.

In the subset selection approach, our goal is to select a non-empty subset of the k populations so that the best population is included in the selected subset with a minimum guaranteed probability P^*. Selection of any subset which includes the best population is called a correct selection (CS). The general approach is to evaluate the infimum of $P(CS|R)$, the probability of a correct selection using the procedure R, over the parameter space $\Omega = \{\underline{\theta}: \underline{\theta} = (\theta_1, \ldots, \theta_k)\}$ and obtain the constants involved in defining R so that

$$\inf_{\Omega} P(CS|R) \geq P^*. \tag{2.1}$$

The condition (2.1) is referred to as the P^*-condition or the basic probability requirement. In order to meet this requirement, one determines the parametric configuration $\underline{\theta}$ for which the infimum in (2.1) is attained. Such a configuration is called a least favorable configuration (LFC). In general, there may not be a unique LFC.

For an extensive survey and bibliography of ranking and selection theory and related topics the reader is referred to the recent book of the authors (1979). Other books in this area are Bechhofer, Kiefer and Sobel (1968), and Gibbons, Olkin and Sobel (1977).

3. Some Subset Selection Procedures.

In this section, we discuss briefly subset selection procedures for normal populations in terms of means, for gamma populations in terms of the scale parameter, and for multivariate normal populations in terms of multiple correlations coefficients. These provide procedures that are applicable in a large number of typical cases.

3.1 Normal Populations. Let $\pi_1, \ldots, \pi_k$ be k independent normal populations with unknown means $\mu_1, \ldots, \mu_k$, respectively, and a common variance σ^2. Let $\overline{X}_i$, $i=1, \ldots, k$, be the sample means based on samples of size n. The best population is the one associated with the largest μ_i. When σ^2 is known, the procedure R_1 proposed by Gupta (1956) selects the population π_i if and only if

$$(3.1) \qquad \overline{X}_i \geq \max(\overline{X}_1, \ldots, \overline{X}_k) - \frac{d_1\sigma}{\sqrt{n}}$$

where $d_1 = d_1(k, P^*) > 0$ is the smallest constant such that the condition (2.1) is satisfied. The LFC is given by $\mu_1 = \ldots = \mu_k$. This implies that d_1 is given by

$$(3.2) \qquad \int_{-\infty}^{\infty} \Phi^{k-1}(x+d_1)\, \phi(x)\, dx = P^*,$$

where $\Phi(x)$ and $\phi(x)$ are the standard normal cdf and density, respectively. The values of d_1 are tabulated for several values of k and P* by Gupta (1963) and Gupta, Nagel and Panchapakesan (1973).

When σ^2 is not known, the procedure R_2 of Gupta (1956) is the same as R_1 with σ replaced by s, where s^2 is the usual pooled estimator of σ^2 based on $\nu = k(n-1)$ degrees of freedom. Here again, the LFC is given by $\mu_1 = \ldots = \mu_k$. The values of the constant d_2 (used in the place of d_1) are tabulated by Gupta and Sobel (1957) for selected values of k, ν, and P*.

The procedures R_1 and R_2 can be modified in the case of the population with the smallest μ_i being defined the best. For procedures involving unequal sample sizes, see Gupta and Huang (1976), and Gupta and Wong (1976).

3.2 Gamma Populations. Let π_i have the associated density

$$(3.3) \qquad f(x, \theta_i) = \begin{cases} \dfrac{x^{r-1}}{\Gamma(r)\theta_i^{\,r}} \exp(-x/\theta_i), & x > 0,\ \theta_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

As we can see, it is assumed that the populations have the same shape parameter $r(> 0)$. Further, r is assumed to be known. Our interest is selecting the population associated with the largest (smallest) θ_i. The gamma distribution not only serves as a model for certain types of measurement, but also includes the case where the observations come from normal populations and the interest is in selecting the population associated with the smallest variance.

For selecting the population associated with the largest θ_i, Gupta (1963) investigated the procedure R_3 which selects π_i if and only if

$$(3.4) \qquad \overline{X}_i \geq b \max(\overline{X}_1, \ldots, \overline{X}_k)$$

where $\overline{X}_1, \ldots, \overline{X}_k$ are means based on samples of equal size n, and the constant b $(0 < b < 1)$ is chosen so that the P*-condition is met. Gupta (1963) has shown that $P(CS|R_3)$ is minimized when $\theta_1 = \ldots = \theta_k$ and that the constant b is given by

$$(3.5) \qquad \int_0^\infty G_\nu^{k-1}(x/b)\, g_\nu(x)\, dx = P^*,$$

where $G_\nu(x)$ is the cdf of a standardized gamma random variable (i.e. with $\theta = 1$) with parameter $\nu/2$ where $\nu = 2nr$. Thus the constant b depends on n and r only through ν and its values are tabulated by Gupta (1963 for selected values of k, P^*, and ν.

For selecting the normal population with the smallest variance, an analogous procedure is given by Gupta and Sobel (1962a) and the appropriate constant can be obtained from the tables in their comparison paper (1962b).

3.3 Multivariate Normal Populations. Let $\pi_1, \ldots, \pi_k$ be k independent p-variate normal population where π_i is $N(\underline{\mu}_i, \Sigma_i)$. Let $\underline{X}_i' = (X_{i1}, X_{i2}, \ldots, X_{ip})$ be a random observation vector from π_i, i=1, ..., p. The populations are ranked in terms of the ρ_i, where ρ_i is the multiple correlation coefficient of X_{i1} with respect to the set $(X_{i2}, \ldots, X_{ip})$. We are interested in selecting a subset containing the population associated with the largest ρ_i. Let R_i denote the sample multiple correlation coefficient between X_{i1} and $(X_{i2}, \ldots, X_{ip})$. Two cases arise: (i) The case in which $X_{i2}, \ldots, X_{ip}$ are fixed, called the conditional case; (ii) The case in which $X_{i2}, \ldots, X_{ip}$ are random, called the unconditional case. *In either* case, Gupta and Panchapakesan (1969) proposed and studied the rule $\mathcal{R}$ which selects π_i if and only if

(3.6) $$R_i^{*2} \geq c \max_{i \leq j \leq k} R_j^{*2}$$

where $R_i^{*2} = R_i^2/(1-R_i^2)$, and $0 < c = c(k, P^*, p, n) < 1$ is chosen to satisfy the P*-requirement. In this case, the infimum of PCS is attained when $\rho_1 = \rho_2 = \ldots = \rho_k = 0$ and the appropriate constant c is given by

(3.7) $$\int_0^{\infty} F_{2q,2m}^{k-1} (x/c)\, f_{2q,2m} (x)\, dx,$$

where $q = \frac{1}{2}(p-1)$, $m = \frac{1}{2}(n-p)$; $F_{r,s}$ denotes the cdf of an F random variable with r and s degrees of freedom, and $f_{r,s}$ denotes the corresponding density.

The values of c are tabulated by Gupta and Panchepakesan (1969) for selected values of k, m, q, and P^*.

4. Selection of Best Predictor Variables.

Many examples of statistical prediction schemes in climatology are available. The prediction is based on a number of predictor variables. While the prediction can be made more accurate by bringing in as many relevant predictor variables as possible, some of them may be highly related among themselves. The problem of selecting the best set of predictor variables arise in different contexts. Stringer (1972 pp. 132-133) has cited examples from literature relating to prediction of precipitation and visibility. Several criteria for defining the best set of predictor variables and various techniques for selecting the best set have been discussed in a nice expository paper by Hocking (1976). Also, a brief review and evaluation of significant methods have been given by Thompson (1978). However, the techniques described by these authors are not designed to find a best set of variables with a guaranteed level of probability. Recently, this problem has been investigated by Arvesen and McCabe (1973, 1975) and Gupta and Huang (1977) under the subset selection formulation described earlier which includes a guaranteed probability of a correct selection. Investigations along these lines continue to be of interest in view of their practical importance.

REFERENCES

Arvesen, J. N. and McCabe, G. P. (1973). Variable selection in regression analysis, Procedings of the University of Kentucky Conference on Regression with a Large Number of Predictors (Ed. W. O. Thompson and F. B. Cady), Dept. of Statist., Univ. of Kentucky, Lexington.

Arvesen, J. N., and McCabe, G. P. (1975). Subset selection problems of variances with applications to regression analysis. J. Amer. Statist. Assoc. 70, 166-170.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist. 25, 16-39.

Bechhofer, R. E., Kiefer, J. and Sobel, M. (1968). Sequential Identification and Ranking Procedures. The University of Chicago Press, Chicago.

Bradley, R. A., Srivastava, S. S., and Lanzdorf, A. (1979). Some approaches to statistical analysis of a weather modification experiment. Comm. Stat.-Theor. Meth., A8, 1049-1082.

Braham, R. R., Jr. (1979). Field experimentation in weather modification. (with comments by R. D. Cook and N. Holschuh, S. M. Dawkins and E. L. Scott, R. D. Elliott, J. A. Flueck, K. R. Gabriel, W. Kruskal, P. W. Mielke, F. Mosteller, J. Neyman, and J. Simpson). J. Amer. Statist. Assoc., 74, 57-104.

Crutcher, H. L. (1969). Uses of some statistics in meterology and climatology. Technical Note No. 88, World Meterological Organization, Geneva, Switzerland, pp. 279-304.

Gibbon, J. D., Olkin, I. and Sobel, M. (1977). Selecting and Ordering Populations. John Wiley, New York.

Glahn, H. R. (1965). Objective weather forecasting by statistical methods. The Statistician, 15, 111-142.

Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Ph.D. Thesis (Mimeo. Ser. No. 150), Institute of Statistics, University of North Carolina, Chapel Hill.

Gupta, S. S. (1963). On a selection and ranking procedure for gamma populations. Ann. Inst. Statist. Math. 14, 199-216.

Gupta, S. S. and Huang, D. Y. (1976a). Selection procedures for the means and variances of normal populations: unequal sample sizes case. Sankhyā Ser. B 38, 112-128.

Gupta, S. S. and Huang, D. Y. (1977e). On selecting an optimal subset of regression variables. Mimeo. Ser. No. 501, Dept. of Statist., Purdue Univ., West Lafayette, Indiana.

Gupta, S. S., Nagel, K. and Panchapakesan, S. (1973). On the order statistics from equally correlated normal random variables. Biometrika 60, 403-413.

Gupta, S. S. and Panchapakesan, S. (1969). Some selection and ranking procedures for multivariate normal populations. Multivariate Analysis - II (Ed. P. R. Krishnaiah), Academic Press, New York, pp. 475-505.

Gupta, S. S. and Panchapakesan, S. (1979). Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations. John Wiley, New York.

Gupta, S. S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. Ann. Math. Statist. 28, 957-967.

Gupta, S. S. and Sobel, M. (1962a). On selecting a subset containing the population with the smallest variance. Biometrika 49, 495-507.

Gupta, S. S. and Sobel, M. (1962b). On the smallest of several correlated F-statistics. Biometrika 49, 509-523.

Gupta, S. S. and Wong, W. Y. (1976b). Subset selection procedures for the means of normal populations with unequal variances: unequal sample sizes case. Mimeo. Ser. No. 473, Dept. of Statist., Purdue Univ., West Lafayette, Indiana.

Hanson, M. A., Barker, L. E., Bach, C. L., Cooley, E. A. and Hunter, C. H. (1979). A bibliography of weather modification experiments. Comm. Stat. - Theor. Meth., A8, 1129-1147.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

Lund, I. A. (1971). An application of stagewise and stepwise regression procedures to a problem of estimating precipitation in California. J. Appl. Meteor., 10, 892-902.

McCutchan, M. H. and Schroeder, M. J. (1973). Classification of meteorological patterns in Southern California by discriminant analysis. J. Appl. Meteor., 12, 571-577.

Mielke, P. W. (1973). Another family of distributions for describing and analyzing precipitation data. J. Appl. Meteor., 12, 275-280.

Mielke, P. W. (1979). Some parametric, nonparametric, and permutation inference procedures resulting from weather modification experiments. Comm. Stat.-Theo. Meth., A8, 1083-1096.

Neyman, J. (1977). Experimentation with weather control and statistical problems generated by it. Applications of Statistics (Ed. P. R. Krishnaiah), North-Holland Publishing Co., Amsterdam, pp. 1-25.

Neyman, J. (1979). Developments in probability and mathematical statistics generated by studies in meteorology and weather modification. Comm. Stat. - Theor. Math., A8, 1097-1110.

Neyman, J., Scott, E. L. and Wells, M. A. (1969). Statistics in meteorology. Rev. Int. Statist. Institute, 37, 119-148.

Stringer, E. T. (1972). Techniques of Climatology, W. H. Freeman Company, San Francisco.

Thompson, M. L. (1978). Selection of variables in multiple regression: Part I. A review and evaluation. Int. Statist. Rev. 46, 1-19.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Mimeograph Series #80-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Some Statistical Techniques for Climatological Data		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER Mimeo. Series #80-1
7. AUTHOR(s) Gupta, S. S. and Panchapakesan, S.		8. CONTRACT OR GRANT NUMBER(s) ONR N00014-75-C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS Purdue University Department of Statistics West Lafayette, IN 47907		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC		12. REPORT DATE January, 1980
		13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Climatological data, ranking and selection, subset selection, normal, gamma, multiple correlations, best predictor variables.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Statistical methods are increasingly being applied in the analysis of climatological data. A brief introduction to subset selection approach in multiple decision theory is given to illustrate the potential applications in climatology.

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)